Crimes in Philadelphia

Team 3

Data Science Capstone Project
Data Acquisition and Pre-Processing Report

Date:

10/20/2020

Team Members:

Name: Raj Patel

Name: Hong Son

Name: Kunal Sharma

[The purpose of this report is to describe the data of your project. It includes three major sections: Data Sources, Data-Processing, and Appendix]

Identifying Data

Data Sources:

[Identify the data sources of your project. It may have more than one data source. Describe each of them and explain why you select the data sources.]

Philadelphia Crime Data:

https://www.kaggle.com/mchirico/philadelphiacrimedata

For our project, we would like to conduct an analysis on a Philadelphia crime dataset published by OpenPhillyData. The dataset dates from 12-31-05 to 03-22-17 (majority of the dates are from 2006-2016). There are 14 total columns in this dataset. They include the district number, sector, dispatch date, dispatch time, location, general code, descriptions of the crime, and the district. This dataset was published to the public so people can assess the crimes happening in Philadelphia.

Weather Data:

http://www.climate.psu.edu/data/city_information/index.php?city=phl&page=dwa&type=big7

To assist in getting more out of our analysis, we will be scraping Philadelphia weather data for Philadelphia from 2007 to 2018. The dataset includes the temperatures in Fahrenheit, the max and min temperature, environment information, wind, and precipitation. We hope to look into the seasonal forecast and see if there are weather trends in crimes.

Acquisition Process:

[Describe the data acquisition process. Is the dataset ready for download? How do you download the data? Do you need to write your own code to acquire the data from a public or private source? Describe how you do it. Are there multiple data sources? How do you integrate the data from multiple sources? Any other process involved in the acquisition process?]

The crime dataset is available as a CSV file while the weather dataset is an XLS file. We don't anticipate having to write code to acquire any data from outside sources. Very little data cleansing had to be done as there was very little missing values and invalid data.

The weather data file will have to be cleansed to include the month and year which are indicated by “Month-Year” on the headers. We will utilize a left join the data using the date field.

Issues:

[Are there any potential issues in data acquisition that have not been solved yet?]

During the weather data cleansing, we are missing a couple of dates that were cut off. Hong has reformat the PDF files so that the dates aren’t cut off and has a new file for the weather data.

We are on track to continue working on the next steps in the project.

Data-Processing:

[Examine the data you have acquired and understand the data properties. Is there any pre-processing you need to do before you can start analyzing the data? For example, missing data, sparsity, noise, veracity, ambiguity, interoperability, etc. Describe each data issue in a sub-section and explain how you clean up the data.]

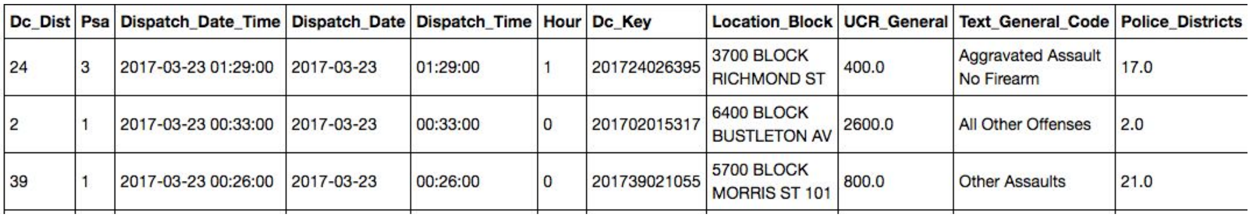
There was missing data in the crime dataset, albeit not a large amount. We removed the rows with missing values. To find locations in North Philadelphia in our data set we first found the Geo location of the town hall ([39.952479, -75.163668]) which is right at the center of philadelphia and then we found the rows in our dataset where the latitude is more than the latitude of the center of Philadelphia. Similarly, for locations in South Philadelphia we found all the rows in our dataset where the latitude is less than the latitude of the center of Philadelphia. To find locations in West Philadelphia we found all the rows where the longitude is less than the longitude of the center of Philadelphia. In this case the rows we found were a combination of locations in north and south philadelphia.

For the weather data cleansing, we didn’t notice any incorrect values. We were only missing a couple dates when we formatted it into PDF files. Hong has reformat the PDF files so that the dates aren’t cut off and has a new file for the weather data.

Appendix

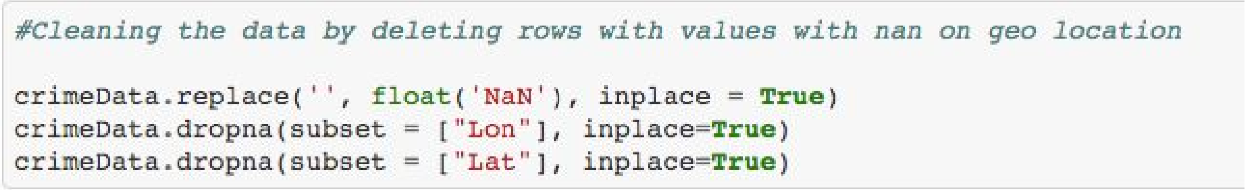
[Provide the code or pseudo code, data definition, sample data, and any other information in the appendix here.]

Crime



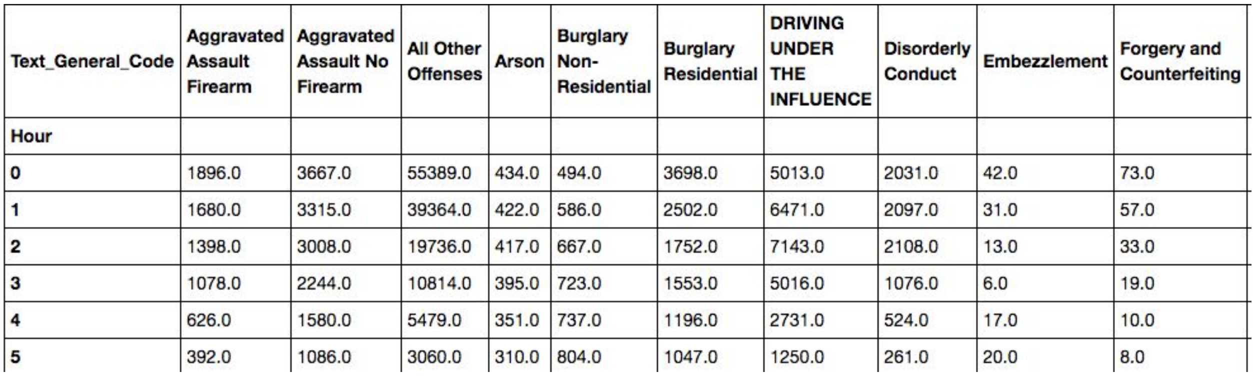
Dc_Dist	Psa	Dispatch_Date_Time	Dispatch_Date	Dispatch_Time	Hour	Dc_Key	Location_Block	UCR_General	Text_General_Code	Police_Districts
24	3	2017-03-23 01:29:00	2017-03-23	01:29:00	1	201724026395	3700 BLOCK RICHMOND ST	400.0	Aggravated Assault No Firearm	17.0
2	1	2017-03-23 00:33:00	2017-03-23	00:33:00	0	201702015317	6400 BLOCK BUSTLETON AV	2600.0	All Other Offenses	2.0
39	1	2017-03-23 00:26:00	2017-03-23	00:26:00	0	201739021055	5700 BLOCK MORRIS ST 101	800.0	Other Assaults	21.0

Top 3 rows in crime data [sorted by Dispatch date]



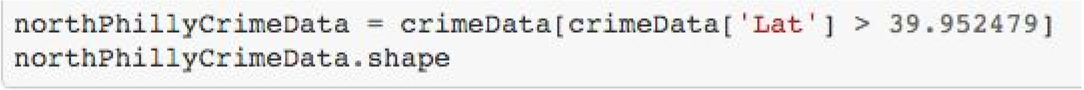
```
#Cleaning the data by deleting rows with values with nan on geo location

crimeData.replace('', float('NaN'), inplace = True)
crimeData.dropna(subset = ["Lon"], inplace=True)
crimeData.dropna(subset = ["Lat"], inplace=True)
```

Remove rows with missing values in locations



Text_General_Code	Aggravated Assault Firearm	Aggravated Assault No Firearm	All Other Offenses	Arson	Burglary Non-Residential	Burglary Residential	DRIVING UNDER THE INFLUENCE	Disorderly Conduct	Embezzlement	Forgery and Counterfeiting
Hour										
0	1896.0	3667.0	55389.0	434.0	494.0	3698.0	5013.0	2031.0	42.0	73.0
1	1680.0	3315.0	39364.0	422.0	586.0	2502.0	6471.0	2097.0	31.0	57.0
2	1398.0	3008.0	19736.0	417.0	667.0	1752.0	7143.0	2108.0	13.0	33.0
3	1078.0	2244.0	10814.0	395.0	723.0	1553.0	5016.0	1076.0	6.0	19.0
4	626.0	1580.0	5479.0	351.0	737.0	1196.0	2731.0	524.0	17.0	10.0
5	392.0	1086.0	3060.0	310.0	804.0	1047.0	1250.0	261.0	20.0	8.0

Creating a dataframe with hours and number of crimes



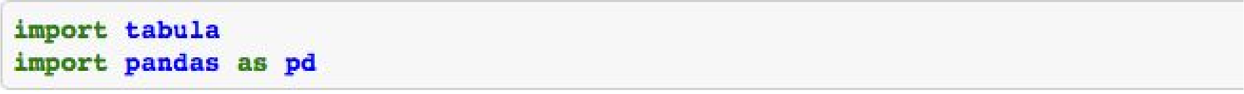
```
northPhillyCrimeData = crimeData[crimeData['Lat'] > 39.952479]
northPhillyCrimeData.shape
```

Looking at regions [North, South, West]

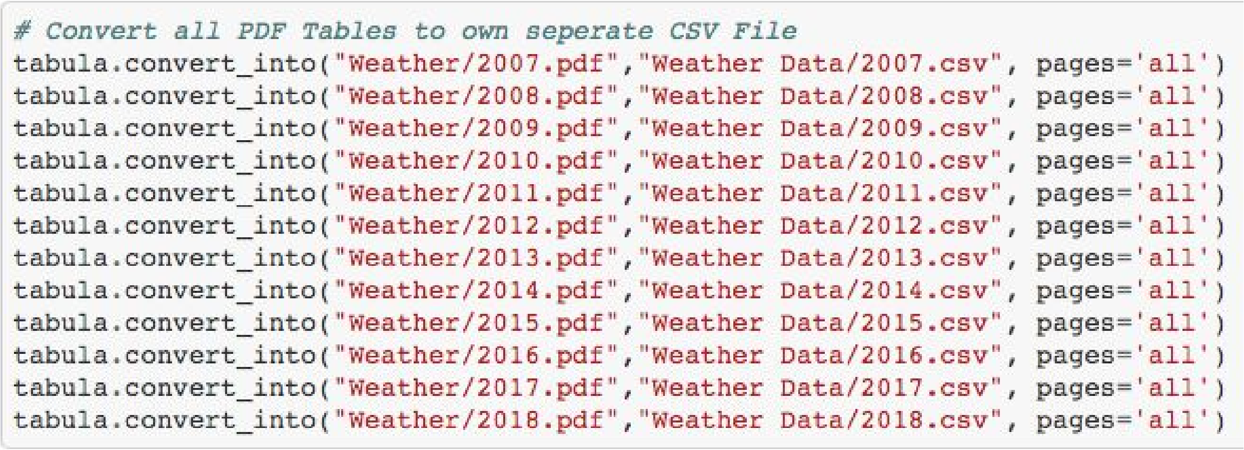
Weather

Weather Data Dictionary -

https://github.com/hongson1218/Crime-in-Philadelphia/blob/master/Weather%20Data%20Dictionary.docx

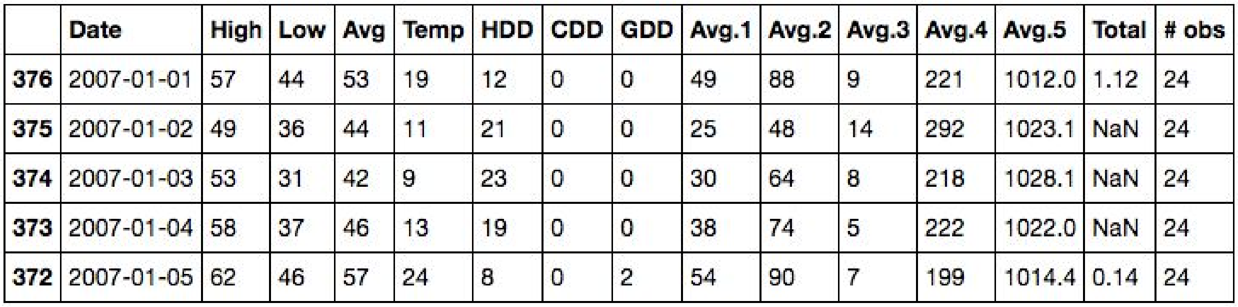


```
import tabula
import pandas as pd
```



```
# Convert all PDF Tables to own seperate CSV File
tabula.convert_into("Weather/2007.pdf","Weather Data/2007.csv", pages='all')
tabula.convert_into("Weather/2008.pdf","Weather Data/2008.csv", pages='all')
tabula.convert_into("Weather/2009.pdf","Weather Data/2009.csv", pages='all')
tabula.convert_into("Weather/2010.pdf","Weather Data/2010.csv", pages='all')
tabula.convert_into("Weather/2011.pdf","Weather Data/2011.csv", pages='all')
tabula.convert_into("Weather/2012.pdf","Weather Data/2012.csv", pages='all')
tabula.convert_into("Weather/2013.pdf","Weather Data/2013.csv", pages='all')
tabula.convert_into("Weather/2014.pdf","Weather Data/2014.csv", pages='all')
tabula.convert_into("Weather/2015.pdf","Weather Data/2015.csv", pages='all')
tabula.convert_into("Weather/2016.pdf","Weather Data/2016.csv", pages='all')
tabula.convert_into("Weather/2017.pdf","Weather Data/2017.csv", pages='all')
tabula.convert_into("Weather/2018.pdf","Weather Data/2018.csv", pages='all')
```

Using tabula to convert pdf to csv files



	Date	High	Low	Avg	Temp	HDD	CDD	GDD	Avg.1	Avg.2	Avg.3	Avg.4	Avg.5	Total	# obs
376	2007-01-01	57	44	53	19	12	0	0	49	88	9	221	1012.0	1.12	24
375	2007-01-02	49	36	44	11	21	0	0	25	48	14	292	1023.1	NaN	24
374	2007-01-03	53	31	42	9	23	0	0	30	64	8	218	1028.1	NaN	24
373	2007-01-04	58	37	46	13	19	0	0	38	74	5	222	1022.0	NaN	24
372	2007-01-05	62	46	57	24	8	0	2	54	90	7	199	1014.4	0.14	24

Looking at top 5 rows

Table of Contributions

The table below identifies contributors to various sections of this document.

	Section	Writing	Editing
1	**Data Sources**	**Hong**	**Raj, Kunal**
2	**Data Pre-Processing**	**Hong**	**Raj, Kunal**
3	**Appendix**	**Hong**	**Raj, Kunal**

Grading

The grade is given on the basis of quality, clarity, presentation, completeness, and writing of each section in the report. This is the grade of the group. Individual grades will be assigned at the end of the term when peer reviews are collected.